Advances in Economics, Business and Management Research, volume 29
International Conference on Innovations in Economic Management and Social Science (IEMSS 2017)

Time Series Analysis: An application of ARIMA model in stock price forecasting
Taking Apple Inc. as an example

Yichen Dong[1, a], Siyi Li[2, b] and Xueqin Gong[3, c]

[1] Economics and Statistics, LSA, University of Michigan Ann Arbor, Ann Arbor, MI 48109, USA

[2]Beijing Normal University, Beijing 100875, China

[3]Economics Lingnan College, Sun. Yet San University, Guangzhou, 512000, China

[a]dongyc@umich.edu, [b]464974001@qq.com, [c]gongxueqin@outlook.com

Keywords: ARIMA model, stock price prediction, time series analysis

Abstract: Time series models have been the foundation of the analysis of a process over a long period of time and their applications are manifold, including sales forecasting, index forecasting etc. In decisions involving uncertainties, time series models are noted as one of the most effective ways of making predictions. Among the many models, the autoregressive integrated moving average (ARIMA) models have been especially popular in time series prediction. This paper presents extensive process of building stock price predictive model using the ARIMA method. Stock data of Apple Inc. obtained from Yahoo! Finance are used. Results revealed that the ARIMA model has a strong potential for short-run forecast and is competitive with other prediction methods in guiding investment decisions.

1. Introduction

Stock price prediction is regarded as one of the most difficult areas to accomplish in financial forecasting. Investors are eagerly looking for forecasting methods that could guarantee easy profit and minimize risks or variations in the stock market. This motivates researchers to develop new predictive models. In the past years, several models had been introduced for stock price prediction. Among them ARIMA models are from statistical viewpoints. They are known for the robustness and efficiency in financial time series forecasting and are extensively used in the research field of economics and finance.

The rest of the paper is organized as follows: Section II presents brief overview of general times series analysis procedures. Section III introduces ARIMA model. Section IV describes the methodology used while section V discusses the experimental results obtained. The paper is concluded in section VI.

2. Time Series Theories

2.1 Time Series

A time series is a set of statistics, usually collected at regular intervals. Time series data occur naturally in many application areas. The aims of time series analysis are to describe and summarize time series data, fit low-dimensional models, and make forecasts.

One way of describing a series is classical decomposition. A series can be decomposed into four elements: Trend — long term movements in the mean; Seasonal effects — cyclical fluctuations related to the calendar; Cycles— other cyclical fluctuations; Residuals — other random or systematic fluctuations. The idea is to create separate models for these four elements and then combine them, either additively ($X_t = T_t + I_t + C_t + E_t$) or multiplicatively ($X_t = T_t \bullet I_t \bullet C_t \bullet E_t$).

2.2 ARIMA Model

If the original process $\{Y_t\}$ is not stationary, we can take first order difference process $X_t = \nabla Y_t = Y_t - Y_{t-1}$ or the second order differences $X_t = \nabla^2 Y_t = \nabla(\nabla Y)_t$ and so on. If we ever find that the

differenced process is a stationary process we can look for a ARMA model of that. The process {Yt} is said to be an autoregressive integrated moving average process, ARIMA(p, d, q), if Xt = ∇dYt is an ARMA(p, q) process.

There are three main steps in modeling data using ARIMA methodology: Model identification, Parameters estimation and Diagnostic checking. The following subsections explain in detail how these steps work.

2.2.a. Model identification:

First thing to do is identify the model by observing the autocorrelation of the time series. Therefore, autocorrelation and partial autocorrelation are needed.

Box-Jenkins method provides a way to identify ARIMA model according to autocorrelation and partial autocorrelation graph of the series. The parameters of ARIMA consist of three components: p (autoregressive parameter), d (number of differencing), and q (moving average parameters).

There are three rules to identify ARIMA model:

1. If ACF (autocorrelation graph) cut off after lag n, PACF (partial autocorrelation graph) dies down, then this process is ARIMA(0, d, n).
2. If ACF dies down, PACF cut off after lag n, the process is ARIMA(n, d, 0).
3. If ACF and PACF die down, a mixed ARIMA model is indicated; further differencing will be needed.

2.2 .b. Parameters estimation:

To estimate the corresponding parameters in the model, simply fit the data to the corresponding ARIMA model, applying ML, and R will run the result. Regarding model selection, we use AICc (Akaike information criterion with a correction). The formula and rule is as follows:

AICc = N*log(SS/N) + 2(p + q + 1)*N/(N – p – q – 2), if no constant term in model

AICc = N*log(SS/N) + 2(p + q + 2)*N/(N – p – q – 3), if constant term in model

N : the number of items after differencing (N = n – d)

SS : sum of squares of differences

p & q : the order of autoregressive and moving average model, respectively

The model with lowest AICc will be selected.

2.2 .c. Diagnostic checking

The diagnose procedure includes observing residual plot and its ACF & PACF diagram, and check Ljung-Box result.

If ACF & PACF of the model residuals show no significant lags, the selected model is appropriate. After ACF/PACF check, Ljung-Box test provides a second way to double check the model. Ljung-Box can test autocorrelation to verify whether the autocorrelations of a time series are different from 0. If the result rejects the hypothesis, data is independent and uncorrelated; otherwise, there remains serial correlation and modification is needed.

3. Methodology

The tool used for implementation is R programming language. Stock data used in this research are published daily stock prices obtained from Yahoo! Finance. In this research the closing price is chosen to represent the price to be predicted. Closing price is chosen because it reflects all the activities of the stock in a trading day.

To determine the best ARIMA model among several experiments performed, the following criteria are used in this study for each stock index.

- Relatively small AIC (Akaike Information Criterion)
- Relatively small standard error
- Relatively high of adjusted R^2

Application

In this section, I will apply the theory to real financial market and analyze the stock price of Apple Inc. (AAPL). The data is selected from Jan 2007 to Dec 2013 with 1761 observations. The line chart

for the close price during that period is shown below

Here we apply the methodology described in section III and testify on the data. The specific steps are as follows:

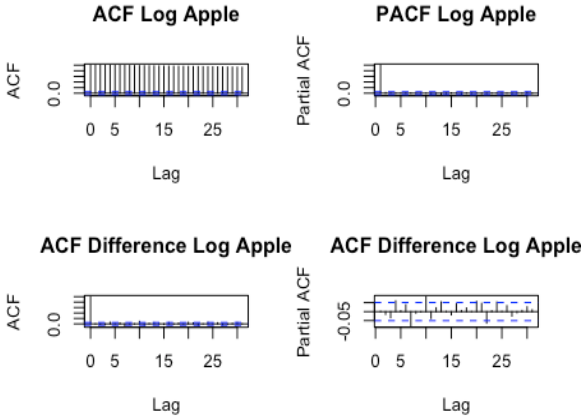
1. Plot ACF/ PACF of the original data
2. If we observe no lags/ no dying down, we'll take difference and plot ACF/ PACF of the differenced data
3. Repeat step 1 &2, until we observe significant lags/ dies down of ACF/ PACF

After coping with data, we apply the rules to identify the model:

If ACF cut off after lag n, PACF dies down: ARIMA(0, d, n)→MA(n)

If ACF dies down, PACF cut off after lag n: ARIMA(n, d, 0)→AR(n)

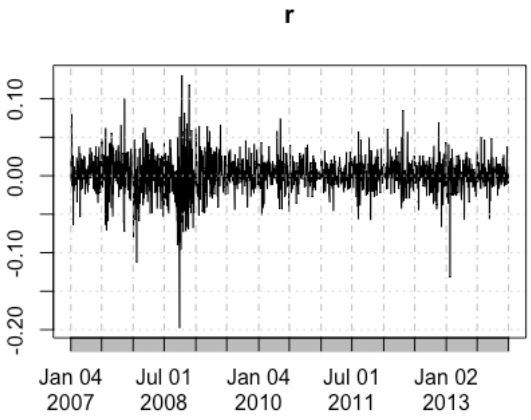
Here I attach the result of AAPL:



As shown in the graph, ACF of Log Apple stock price slowly decreases (not dies down), meaning the model needs differencing.
PACF of Log Apple shows significant value at lag 1 and then PACF cuts off. The model for Log Apple stock price might be ARIMA(1,0,0)
After taking one step difference, we observe that ACF of differences of log Apple with no significant lags; PACF of differences of log Apple reflect no significant lags. The model for differenced log Apple series is thus a white noise, and the original model resembles random walk model ARIMA(0,1,0)

The above procedures also make sense in financial perspective besides making the data stationary. To analyze the stock price, we usually calculate the logged return of the stock.

The following plot shows the daily logged return of AAPL.



From the above plot, it is induced that no trend or seasonality presents for this time series with a mean 0. Further steps are need to testify this rough induction.

Trend

To test whether there is a drift or a trend for the return, I apply the Augmented Dickey–*Fuller* (ADF) test. The model for the ADF test is:

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \cdots + \delta_{p-1} \Delta X_{t-p-1} + \varepsilon_t$$

The result is shown below:

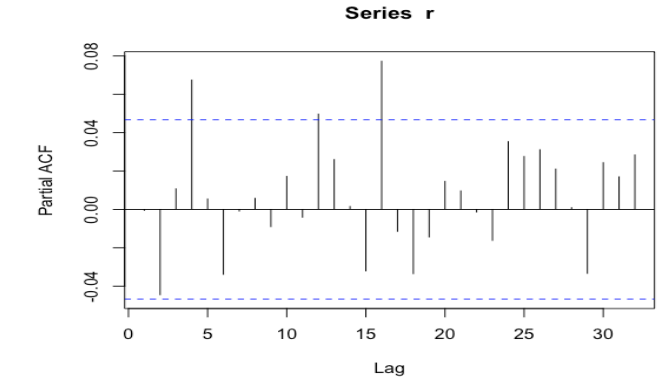
```
Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
      Min        1Q    Median        3Q       Max
-0.197992 -0.010895  0.000486  0.012171  0.128797

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.681e-03  1.110e-03   1.514   0.1301
z.lag.1     -1.049e+00  3.401e-02 -30.845   <2e-16 ***
tt          -6.091e-07  1.085e-06  -0.561   0.5746
z.diff.lag   4.298e-02  2.398e-02   1.792   0.0733 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02273 on 1736 degrees of freedom
Multiple R-squared:  0.5037,    Adjusted R-squared:  0.5029
F-statistic: 587.4 on 3 and 1736 DF,  p-value: < 2.2e-16


Value of test-statistic is: -30.8449 317.1356 475.7034

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

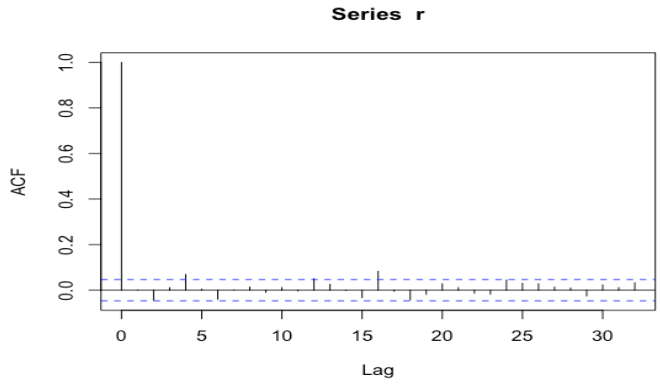
Result shows that the intercept is not significant. We cannot reject the hypothesis that the intercept is 0, in other words, there is no drift. Also, there is no linear trend for this time series, since the coefficient for tt is not significant.

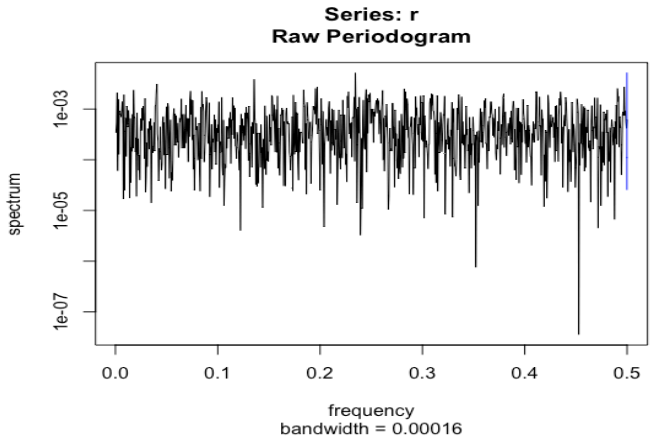
Additionally, this test shows no unit root present since the null hypothesis of $\gamma = 0$ is rejected, indicating the model is not a random walk. We also notice that γ =-1.056, whose absolute value is greater than 0. The AR part of the model is stationary.

Seasonality

By viewing the ACF, PACF, and spectrum Periodgram which are shown below, we cannot find an evidence for seasonality.

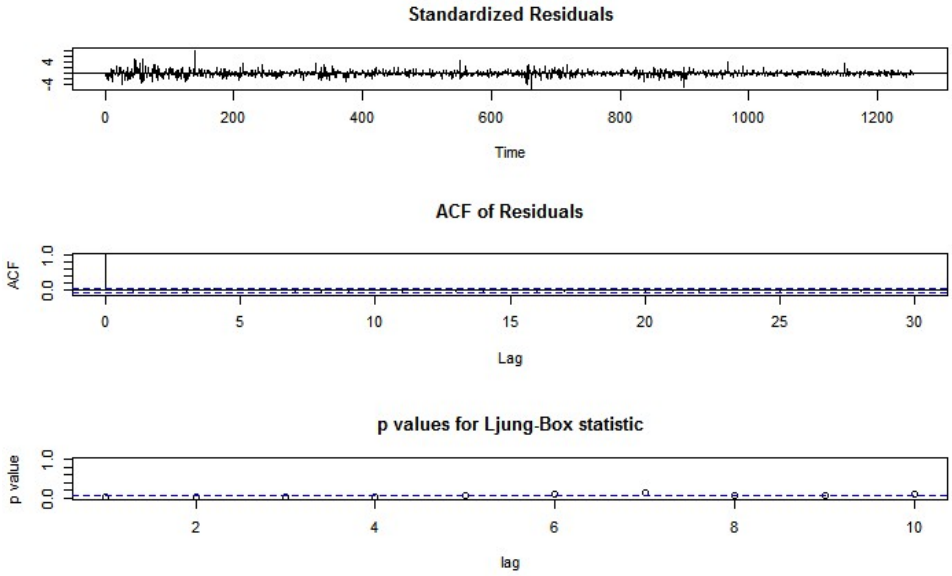






The I first tried ARIMA(0,1,1). The regression result and diagnostic of residuals are summarized in the following. From the ACF plot and the Ljung-Box statistics, we can see that the residuals are almost uncorrelated.

```
> fit1 = arima(r, order=c(0,1,1))
> summary(fit1)
Series: r
ARIMA(0,1,1)

Coefficients:
          ma1
      -1.0000
s.e.   0.0042

sigma^2 estimated as 0.0004277:  log likelihood=3088.16
AIC=-6172.33   AICc=-6172.32   BIC=-6162.06

Training set error measures:
                      ME       RMSE        MAE MPE MAPE     MASE
Training set -0.0007476267 0.02067255 0.01441089 NaN  Inf 1.000537
```

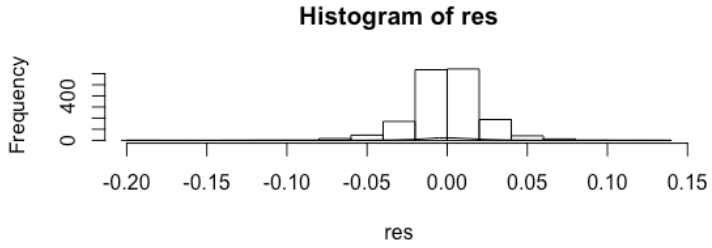


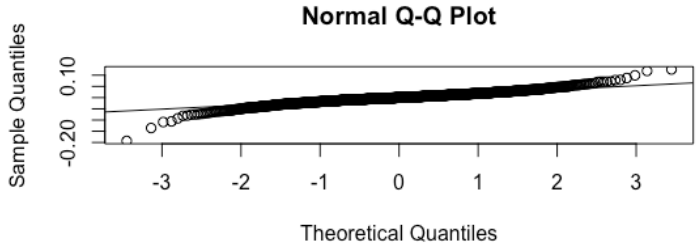
To testify the model, I use the build-in function auto arima in R to determine a more appropriate model. The result shows that a ARIMA(4,0,3) is selected

```
Series: r
ARIMA(4,0,3) with non-zero mean

Coefficients:
         ar1      ar2     ar3     ar4      ma1     ma2      ma3  intercept
      0.6040  -0.8765  0.5489  0.0203  -0.6047  0.8494  -0.5281     0.0011
s.e.  1.6317   0.0712  1.5011  0.0271   1.6339  0.0695   1.4404     0.0006

sigma^2 estimated as 0.0005164:  log likelihood=4169.39
AIC=-8320.78   AICc=-8320.68   BIC=-8271.52

Training set error measures:
                       ME       RMSE       MAE MPE MAPE      MASE          ACF1
Training set -1.419408e-06 0.02267282 0.0160822 Inf  Inf 0.7035221 -0.0003520802
```

Model selection

The AICc for selecting the models has been introduced in the previous section and now let us put it into practice. Comparing ARIMA(0,1,1) and ARIMA(4,0,3), we observe a lower value of AICc(-8320.68) for ARIMA(4,0,3) than ARIMA(0,1,1) (-6172.32). Hence we select ARIMA(4,0,3) as our fitted model to make further predictions. Besides that, we can also check the normality of the residuals. From the histogram and qq-plot of the residual, we can see that the residuals are not normal distributed. The flat density curve and the invert S-shaped qq-plot indicate that the density of the residual should be fat tailed.





Forecasting

Using the ARIMA(4,0,3) model with no drift to forecast the logged return, I get the following result.

```
Forecast method: ARIMA(4,0,3) with non-zero mean

Model Information:
Series: r
ARIMA(4,0,3) with non-zero mean

Coefficients:
         ar1      ar2     ar3     ar4      ma1     ma2      ma3  intercept
      0.6040  -0.8765  0.5489  0.0203  -0.6047  0.8494  -0.5281     0.0011
s.e.  1.6317   0.0712  1.5011  0.0271   1.6339  0.0695   1.4404     0.0006

sigma^2 estimated as 0.0005164:  log likelihood=4169.39
AIC=-8320.78   AICc=-8320.68   BIC=-8271.52

Error measures:
                       ME       RMSE       MAE MPE MAPE      MASE          ACF1
Training set -1.419408e-06 0.02267282 0.0160822 Inf  Inf 0.7035221 -0.0003520802

Forecasts:
     Point Forecast       Lo 80      Hi 80       Lo 95      Hi 95
1762  -0.0000626696 -0.02918528 0.02905994 -0.04460186 0.04447652
1763   0.0013773567 -0.02774526 0.03049997 -0.04316184 0.04591656
1764   0.0017809791 -0.02735265 0.03091461 -0.04277506 0.04633702
1765   0.0008303380 -0.02830364 0.02996432 -0.04372624 0.04538691
1766   0.0004553854 -0.02871062 0.02962140 -0.04415018 0.04506095
1767   0.0013128992 -0.02785427 0.03048007 -0.04329444 0.04592024
1768   0.0016458310 -0.02753775 0.03082941 -0.04298661 0.04627827
1769   0.0008702193 -0.02831344 0.03005388 -0.04376233 0.04550277
1770   0.0005730484 -0.02862734 0.02977344 -0.04408509 0.04523119
1771   0.0012735184 -0.02792713 0.03047417 -0.04338502 0.04593206
```

4. Conclusions

I fit the logged return of AAPL data in a ARIMA(4,0,3) model with no drift:

$Yt = 0.6040Yt-1 – 0.8765Yt-2 + 0.5489Yt-3 +0.0203 Yt-4 - 0.6047\varepsilon t-1 + 0.8494\varepsilon t-2 -0.5281\varepsilon t-3+ \varepsilon t$

and the 10-step predicted logged return values are listed in the above forecasts summary.

References

[1] Sandra Lach Arlinghaus, PHB Practical Handbook of Curve Fitting. CRC Press, 1994.

[2] William M. Kolb. Curve Fitting for Programmable Calculators. Syntec, Incorporated, 1984.

[3] S.S. Halli, K.V. Rao. 1992. Advanced Techniques of Population Analysis. ISBN 0306439972 Page 165 (cf. ... functions are fulfilled if we have a good to moderate fit for the observed data.)

[4] Numerical Methods in Engineering with Python 3. By Jaan Kiusalaas. Page 21.

[5] Fitting Models to Biological Data Using Linear and Nonlinear Regression. By Harvey Motulsky, Arthur Christopoulos.

[6] *Imdadullah.* "Time Series Analysis". *Basic Statistics and Data Analysis. itfeature.com*. Retrieved 2 January 2014.

[7] *Bloomfield, P. (1976). Fourier analysis of time series: An introduction. New York: Wiley.* ISBN 0471082562.

[8] *Lin, Jessica; Keogh, Eamonn; Lonardi, Stefano; Chiu, Bill (2003). "A symbolic representation of time series, with implications for streaming algorithms". Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. New York: ACM Press.* doi:10.1145/882082.882086.